LAMP-TR-002
CFAR-TR-841
CS-TR-3697

October 1996

The Function of Documents

David Doermann, Ehud Rivlin, Azriel Rosenfeld

Language and Media Processing Labratory
Instititue for Advanced Computer Studies
College Park, MD 20742

Abstract

The purpose of a document is to facilitate the transfer of information from its author to its readers. It is the author's job to design the document so that the information it contains can be interpreted accurately and efficiently. To do this, the author can make use of a set of stylistic tools. In this paper we introduce the concept of document functionality, which attempts to describe the roles of documents and their components in the process of transferring information. A functional description of a document provides insight into the type of the document, into its intended uses, and into strategies for automatic document interpretation and retrieval. To demonstrate these ideas, we define a taxonomy of functional document components and show how functional descriptions can be used to reverse-engineer the intentions of the author, to navigate in document space, and to provide important contextual information to aid in interpretation.

***The support of the LAMP Technical Report Series and the partial support of this research by the National Science Foundation under grant EIA0130422 and the Department of Defense under contract MDA9049-C6-1250 is gratefully acknowledged.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
OCT 1996		**00-10-1996 to 00-10-1996**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
The Function of Documents	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Language and Media Processing Laboratory,Institute for Advanced Computer Studies,University of Maryland,College Park,MD,20742-3275**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**27**	

1 Documents as Message Conveyors

Written documents have long been the preferred medium for the transfer of information across both time and space. In this sense, the general purpose or "function" of a document is to store data produced by a sender in a symbolic form to facilitate transfer to a receiver. Traditionally, the data takes the form of a set of markings on a page, with the sender corresponding to the "author", and the receiver to the "reader". In this paper we limit ourselves to the understanding and interpretation of these "traditional" 2D documents which the reader receives visually. We do not consider 3D artifacts that might be used to transfer information (not even cases such as Braille, bas-relief, etc., which are nearly 2D), nor do we treat time-varying "documents" such as audio or video, though it seems clear to us that our approach could be extended to such non-traditional domains.

When documents are regarded as message conveyers, we can classify them according to the type of message that is conveyed. We will differentiate between three classes: informational, instructional, and identificational.

- **Informational:** The message can contain "expository" information such as might be found in a report, dictionary, newspaper, novel, catalogue or the like.
- **Instructional**: The message may have an instructional content, relating to an action or series of actions, such as found in a recipe book, a do-it-yourself manual, a how-to-get-there description, a road sign, etc. A special case of this category, which we shall refer to as the "dialogue" sub-category, involves instructions about changing the document itself. This might, for example, involve the intentional placement of additional markings in the original page, as in filling out a form. "Dialogue" documents include diaries, postcards, tax forms, and bank checks, for example.
- **Identificational:** In this class the message is intended to identify a location (a street sign), an object (a car license plate), or a person (a name tag), for example. This class of documents usually has a locational component, so that the nature of the transferred information depends on the location of the document. A street sign taken away from its proper place conveys deceptive information.

In any of these classes of documents, the message can be represented in various ways. Media that can be used to convey messages include text, graphics, and images. The message may be representational, as in the case of an image, map or diagram which has some isomorphic relationship with the real world, or it may be represented by arbitrary symbols like those of a modern alphabet. Pictograms are an intermediate type of representation. A narrative uses words to represent a spatio-temporal structure; a (static) image or a map can represent only spatial relations. We often, of course, use mixed representations.

In addition to its message, a document can be evaluated with respect to its esthetics. One can evaluate the whiteness of the page and the sharpness of the markings, the shapes of the symbols (calligraphy), the beauty of a painting or a poem, and so on. In this paper, however, we will emphasize the type of message that the document is intended to convey.

The types of messages describe above were formulated from the author's point of view. The reader, the receiver of the document, may have different goals, and may abstract the document's contents at many different levels. Readers can become quite skilled at abstracting task-dependent information from a document and using this information to establish a context for further interpretation. For example, when looking for documents created on a specific date, an experienced reader can rapidly locate the dates of documents

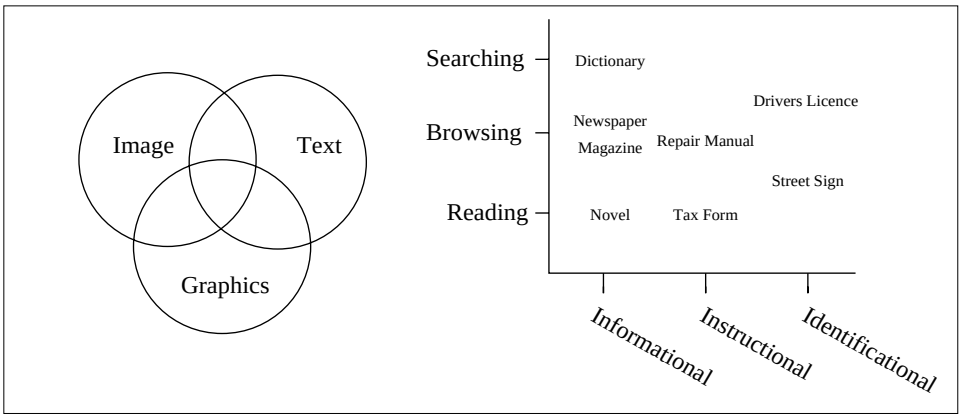


Figure 1: Classification: A document can be represented using any combination of the three media: Images, graphics, and text. The other two dimensions are the type of message that is conveyed and the way the reader interacts with the document.

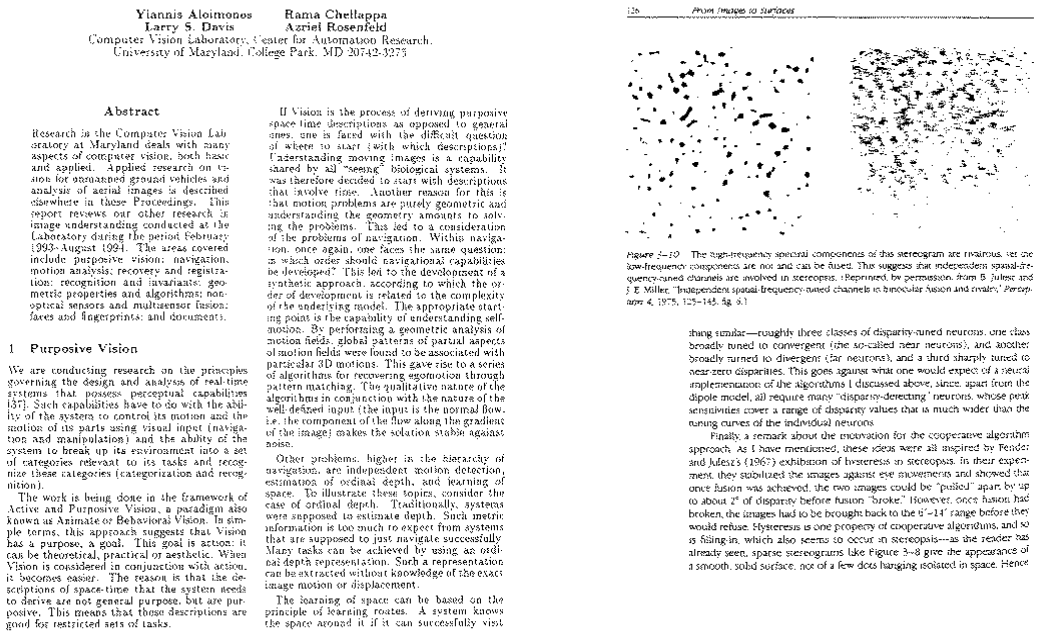
such as business letters and forms without reading them entirely. If it is then decided to "read" the document, the context helps with its correct interpretation and provides a framework in which to proceed through it in an orderly fashion. We can distinguish three basic ways of doing this:

- **Reading** - which usually involves examining the document from beginning to end. This mode is ordinarily used for letters, articles, and many types of books. The examination may be more or less thorough, ranging from proofreading to skimming.
- **Browsing** - which involves examining only selected parts of the document to determine if more in-depth examination of these parts is required. This mode is ordinarily used for newspapers, magazines, and journals.
- **Searching** (or referencing) - which involves looking for a specific piece of information in the document. This mode is ordinarily used for reference books such as dictionaries, encyclopedias, directories, manuals, handbooks, catalogs, etc.

As Figure 1 shows, the mode of transfer of the information and the type of message are relatively independent. Examples of each of the modes are shown in Figures 2 - 4.

These modes of interaction with a document apply not only to text-intensive documents; they can also apply to documents which are primarily representational, such as maps and drawings. However, the processes used to read, browse, or search a document depend on the document type. For example, browsing a newspaper and browsing a map have the same basic goal of examining only selected parts, but the methods which are used to accomplish this are quite different. Similarly, searching a phone book and searching a map both require "navigating" and making decisions based on partial information, but they involve different processes. For phone books, one uses index terms and alphabetical relationships; for maps, one uses symbols or landmarks and spatial relationships.

Although a particular document may be designed primarily for a particular mode of transfer, it may also be used in other ways. A recipe, for example, may be primarily instructional and we read it to follow the step by step procedure. We may, however, have a collection of recipes in a cookbook, and browse it to look for something to make, or perhaps search it to find a particular recipe; both of these are informational functions.



Maryland Progress in Image Understanding

Yiannis Aloimonos Rama Chellappa
Larry S. Davis Azriel Rosenfeld
Computer Vision Laboratory, Center for Automation Research,
University of Maryland, College Park, MD 20742-3275

Abstract

Research in the Computer Vision Laboratory at Maryland deals with many aspects of computer vision, both basic and applied. Applied research on vision for unmanned ground vehicles and analysis of aerial images is described elsewhere in these Proceedings. This report reviews our other research in image understanding conducted at the Laboratory during the period February 1993–August 1994. The areas covered include purposive vision; navigation; motion analysis; recovery and registration; recognition and invariants; geometric properties and algorithms; non-optical sensors and multisensor fusion; faces and fingerprints; and documents.

1 Purposive Vision

We are conducting research on the principles governing the design and analysis of real-time systems that possess perceptual capabilities [37]. Such capabilities have to do with the ability of the system to control its motion and the motion of its parts using visual input (navigation and manipulation) and the ability of the system to break up its environment into a set of categories relevant to its tasks and recognize these categories (categorization and recognition).

The work is being done in the framework of Active and Purposive Vision, a paradigm also known as Animate or Behavioral Vision. In simple terms, this approach suggests that Vision has a purpose, a goal. This goal is action: it can be theoretical, practical or aesthetic. When Vision is considered in conjunction with action, it becomes easier. The reason is that the descriptions of space-time that the system needs to derive are not general purpose, but are purposive. This means that these descriptions are good for restricted sets of tasks.

If Vision is the process of deriving purposive space time descriptions as opposed to general ones, one is faced with the difficult question of where to start (with which descriptions)? Understanding moving images is a capability shared by all "seeing" biological systems. It was therefore decided to start with descriptions that involve time. Another reason for this is that motion problems are purely geometric and understanding the geometry amounts to solving the problems. This led to a consideration of the problems of navigation. Within navigation, once again, one faces the same question: in which order should navigational capabilities be developed? This led to the development of a synthetic approach, according to which the order of development is related to the complexity of the underlying model. The appropriate starting point is the capability of understanding self-motion. By performing a geometric analysis of motion fields, global patterns of partial aspects of motion fields were found to be associated with particular 3D motions. This gave rise to a series of algorithms for recovering egomotion through pattern matching. The qualitative nature of the algorithms in conjunction with the nature of the well-defined input (the input is the normal flow, i.e. the component of the flow along the gradient of the image) makes the solution stable against noise.

Other problems, higher in the hierarchy of navigation, are independent motion detection, estimation of ordinal depth, and learning of space. To illustrate these topics, consider the case of ordinal depth. Traditionally, systems were supposed to estimate depth. Such metric information is too much to expect from systems that are supposed to just navigate successfully. Many tasks can be achieved by using an ordinal depth representation. Such a representation can be extracted without knowledge of the exact image motion or displacement.

The learning of space can be based on the principle of learning routes. A system knows the space around it if it can successfully visit

9

Figure 2: Reading documents

Figure 3: Browsing documents

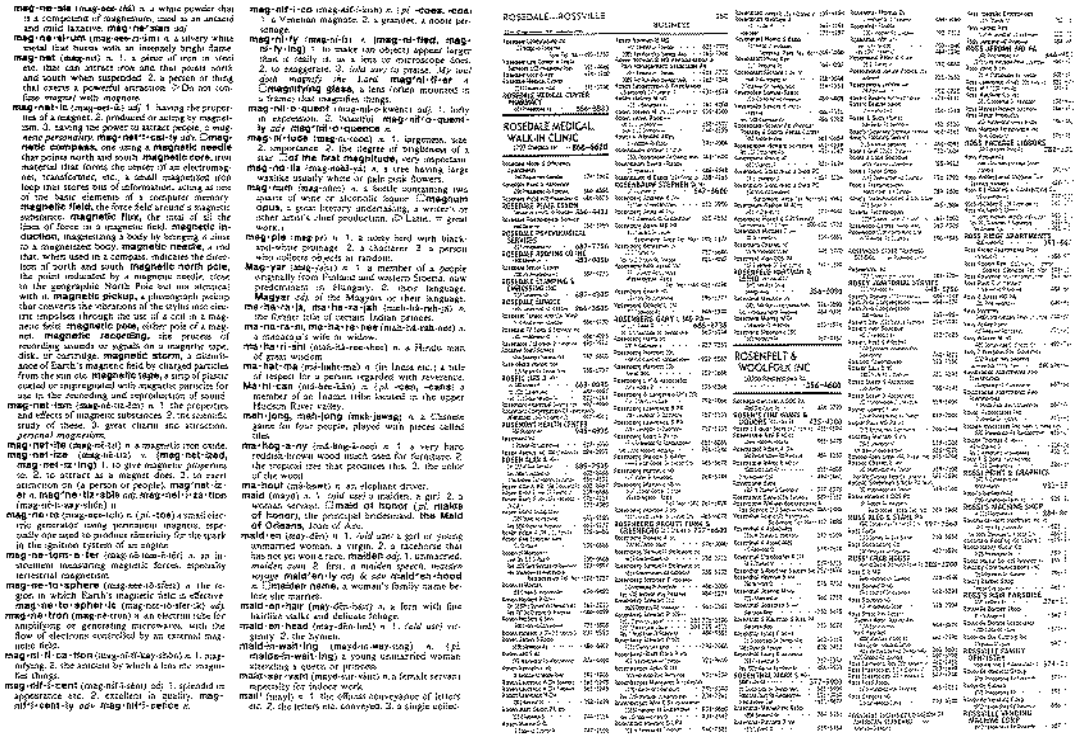


Figure 4: Searching documents

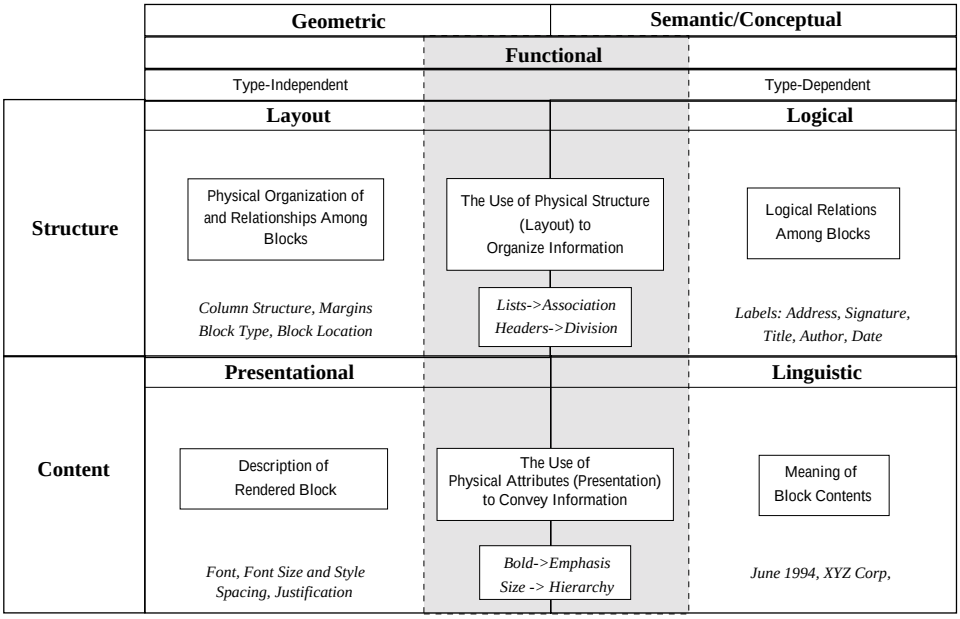


Figure 5: The relationship of geometric, semantic and functional descriptions.

A great deal of work has been done on the analysis of document structure. Almost all of this work, however, has involved models for specific classes of documents. We believe that significant progress in the automated analysis of general classes of documents depends on the development of a general framework for describing document structure. This paper attempts to develop a such a framework.

2 Document Structure

In this section, we first consider traditional views of document organization and show how a document's functional organization (i.e. organization in information transfer terms) is related to its geometric and semantic organizations (Section 2.1). We then illustrate how the author and the reader are able to use the design of a document to impose functional organization on the document (Section 2.2). Finally, in Section 2.3 we make an analogy between the components of a document, which is a device for transferring information, and the parts of a tool, which is a device for transferring force.

2.1 Levels of Document Organization

In document understanding, documents have traditionally been viewed according to their geometric and semantic organizations, as shown in Figure 5[1]. Both organizations have a common *content* which represents a base level of data (typically text, but also possibly including graphics or images). The content's *geometric* nature refers to how it is presented on the page (for example, typeface and font size, for text; line widths and symbols, for graphics), and its *semantic* nature refers to its meaning.

Similarly, a document has both geometric and semantic *structure*. The *layout* structure corresponds to the organization of the document into geometric groupings such as charac-

[1] This is the view taken in the ODA standard [5].

Structure	Example	Use
header	centered	relative importance, focal point
list	enumerated itemized	conveys temporal sequence suggests similar level of descriptiveness
separator	white space or rule line	physical and possibly semantic dis-association
attachment	footnote boxed text sidebar	supplemental information under some semantic hierarchy
illustration	table figure	supplemental information - preserves 2D associations graphical representation of information

Figure 6: Some structures and their uses

ters, lines, blocks, columns, etc. It describes the relationships among these components and the relationships of the individual components to the entire page. The *logical* structure, on the other hand, organizes the content according to the interpretation of the reader, and also provides global relationships such as reading order. The logical structure corresponds to the document's semantic or conceptual organization.

We claim that there is a level of document organization, which can be regarded as intermediate between the geometric and semantic levels, that relates to the efficiency with which the document transfers its information to the reader. We refer to this level as the *functional* level.

A document obeys conventions such as the use of an alphabet and a language common to the author and reader, and the use of standard presentation rules such as word and line spacing, punctuation, etc. As the information content of the document becomes more complex, these conventions may no longer be adequate for efficient information transfer. Appropriate structures can be used to enhance efficient transfer of information and reduce its ambiguity. For example, an author may use page or section headers to "summarize" content; ordered lists to enumerate or itemize information; separators to "punctuate"; attachments (such as footnotes and sidebars) to subordinate; tables or graphs to present numeric data; maps to present spatial data and their interrelationships. (Note that graphs and maps involve augmenting the basic language with more expressive constructs.) Figure 6 shows some examples of such structures.

As an illustration of the relationship between the geometric, functional, and semantic organizations of a document, consider a **block** of text at the top of a page. Its dimensions and location on the page, as well as properties of its components, are geometric or layout attributes. The fact that we have grouped the components together to form the block is based on geometric proximity. We can use the block's attributes (position, size, etc.) in a class-independent manner to conclude that the block is a **header**; this describes it functionally. If we make a class-dependent identification of the block as a **title**, we have given it a semantic description. Note that a similar block could be a running head or a letterhead in a different context.

The functional description of a document is often independent of document type and can be derived from geometric considerations. Headers, footers, lists, tables, and graphics are examples of generic structures which can be common to many types of documents. Such

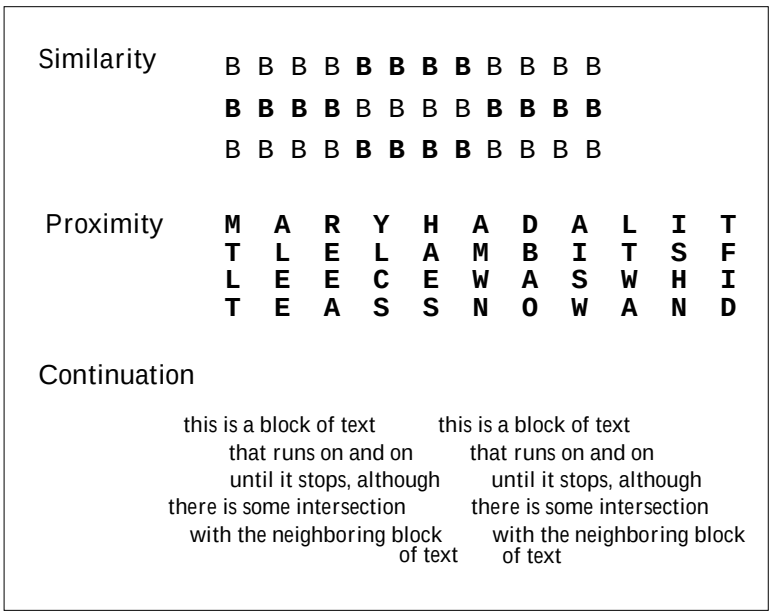


Figure 7: Document interpretation is consistent with the principles of Gestalt.

functional structures will be referred to as class-independent.

If the type of the document is known (for example, business letters or memos, forms, advertisements, or technical articles), a component can have functionality with respect to the documents of that type. For example, in a letter, functional components may include the sender, receiver, date, and salutation. Such functional components will be referred to as class-dependent. The formats used in documents of specific types, such as business letters or journal articles, also serve to enhance information transfer by helping to organize and prioritize the information.

2.2 Functional Document Design

Because the transfer of information to the reader of a document is done using vision as a medium, documents should be designed in accordance with basic perceptual principles such as the principles of Gestalt [6]. When we use white spaces as separators, the principle of proximity, which states that elements which are closer together tend to be grouped together, is being applied. According to this principle, the space between lines should be greater than the average space between words and letters. The principle of good continuation, according to which elements that lie along a common line or smooth curve are grouped together, causes the white spaces that border a column to be seen as units, thus separating the column from its neighbors. The principle of similarity, which states that elements that are similar in physical attributes, such as color, orientation, or size, are grouped together, causes words in boldface to group together. Figure 7 shows some examples of the operation of Gestalt laws [6].

The author of a document can take advantage of these principles to design the document so that the reader can use it effectively. Authors typically use combinations of layout and emphasis to convey an intended organization, or to assign priorities to specific components.

Within a document, structures such as those shown in Figure 6 can be used as aids in the organization of information. A list, for example, suggests a meaningful temporal or

set relationship between its items. A figure and the corresponding caption are interpreted as an illustration of some concept or fact in the text. Higher-level constructs such as sections/subsections, columns, indices, or running heads aid in organizing a document at a more global level.

Other techniques can be used to attract (or suppress) a reader's attention. At a page level, an author can use headers and increase their point size, use all caps, and/or center them to make them more prominent. At a word or phrase level, the author can use bold-face or italic fonts in a similar way to draw attention. Text which is seen as unimportant can be put in "fine print" with opposite results.

As Figure 8 illustrates, documents can be designed to allow the derivation of plausible organizational structures in the absence of class models, even when the meaning of the document is not understood.

2.3 Informational Advantage

Much of the work on function-based object recognition [9, 10, 13, 14] has dealt with cases in which the object functions as a "tool". A tool [3] is an object that receives input force from a "source" and delivers output force to a "receptor". In this general sense, a chair can be regarded as a primitive "tool": it receives the weight of the sitter's body at its "input" end (the seat) and delivers it to the output end(s) (the legs or base on which it rests on the floor), thus allowing the floor to support the sitter at the height of the seat. Similarly for a cup, which can contain liquids; a knife, which can be used to cut; and so on.

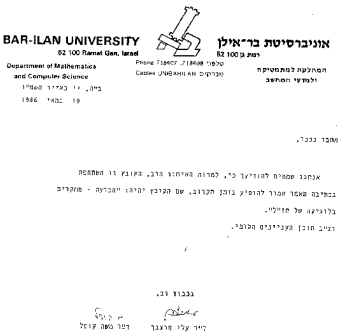
A document is a message conveyer, an object which transfers information. Just as a function of an object such as a tool can be associated with the type of force it transfers, and how well or efficiently it does so (a well-designed tool will transfer force efficiently), a function of a document can be associated with the type of information it transfers ("informational" (i.e. expository), instructional, or identificational) and how well or efficiently it does so.

When we analyze the functionality of a tool we try to recognize its functional parts [9]. A lever has an input end and an output end; the first should facilitate grasping, the second should facilitate application of force (torque). The lever amplifies the torque applied to it by its user, and constitutes a primitive tool (a "simple machine"). In the tool recognition process we try to establish a mapping between shape parts and functional parts [9]. We can take a similar approach in the document domain and define functional parts which play roles in the information transfer process. These functional parts of a document will be called *information units*.

An information unit is the base level of representation necessary for the reader to perform some task involving the transfer of information. For example, if the task is to recognize individual characters, the information unit is typically a single symbol. If the task involves searching a phone book, the information unit may be a single listing; if the task is to read a book, the information unit may be a block of text which corresponds to a paragraph or section.

The analog of a tool in the document domain is an *information structure*. This is a document component consisting of one or more information units - for example, a list or table.

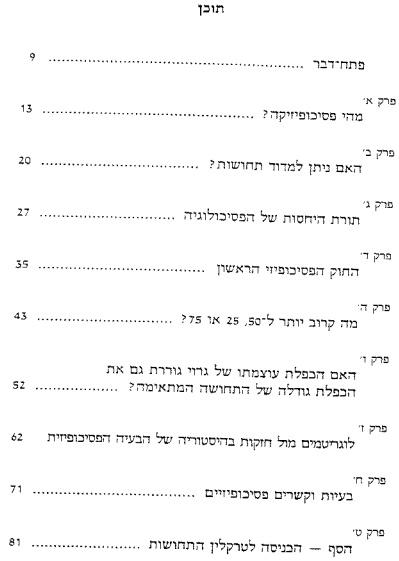
For a tool we define the *mechanical advantage* as the ratio of the output force to the input force. In a hammer, for example, this ratio is high because of the long handle (as well as the concentration of mass in the hammerhead). Thus the geometry of a tool contributes



BAR-ILAN UNIVERSITY
52 100 Ramat Gan, Israel
Department of Mathematics and Computer Science

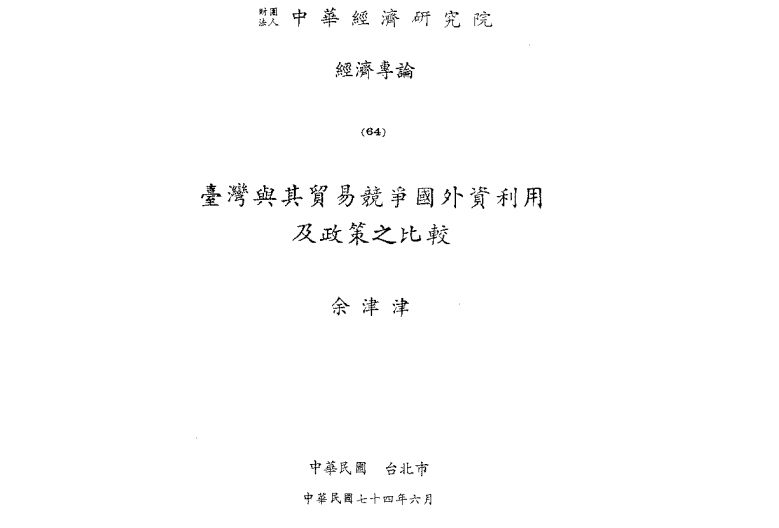
אוניברסיטת בר־אילן
רמת גן 52 100
המחלקה למתמטיקה ולמדעי המחשב

(a)



תוכן

(b)



財團法人 中華經濟研究院

經濟專論

(64)

臺灣與其貿易競爭國外資利用及政策之比較

余津津

中華民國 台北市

中華民國七十四年六月

(c)

Figure 8: Recognizable structure without content

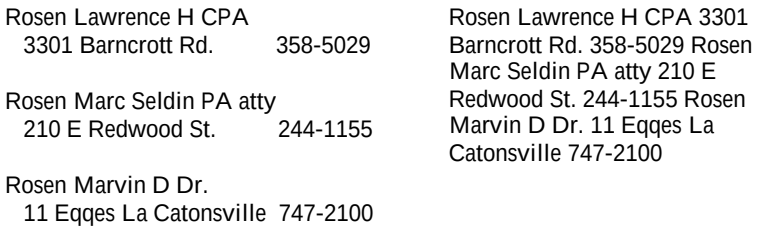


Figure 9: Proper design achieves an information advantage: A list as an "information machine"

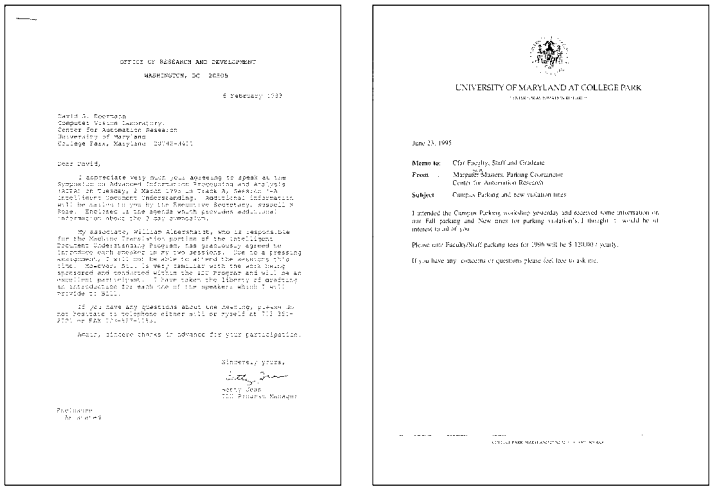
to its mechanical advantage. In a similar manner we expect a well-designed document to transfer information efficiently and to give some *informational advantage*. It is evident that proper document design achieves such an advantage; a well-designed text can be read (or browsed, or searched) much more rapidly than an unstructured text, as illustrated in Figure 9 (see also Figure 7).

3 Exploiting Function

In order to effectively process a document, most document image understanding systems rely on relatively specific information about a restricted domain in order to accurately model the expected document class(es). This allows the system to richly interpret the document, and extract detailed information about its content. For example, in the domain of business letters, a great deal of work has been done on both their structural and logical interpretation ([1], [2], [4], [7], [8], [15], [16]). Unfortunately, for less homogeneous environments this approach cannot be effectively applied. As the set or stream of documents becomes more diverse (both intra-class and inter-class), the formulation of models becomes more difficult. Functional interpretation of documents can greatly facilitate tasks associated with their classification and use. In the following paragraphs we give three examples of tasks which can be addressed by identifying functionally meaningful constructs in documents.

Use Classification: In Section 1, we identified three major ways in which a reader can use a document: reading, browsing, and searching. Documents designed for these purposes can be grossly characterized by the size and organization of their information units, which can be identified by repetitive patterns in the document. For example, reading documents such as journal articles tend to have a single read-order and large information units; browsing documents, such as newspapers or popular magazines, tend to have multiple head-body structures, since their designer's goal is to give the reader quick access to the contents with "handles"; and searching documents tend to have many small information units such as the entries in an index or phone book. An instructional document intended for modification by the reader, such as a form, is characterized by small, blank information units such as horizontal line segments or boxes (including small check boxes). We will demonstrate this approach to document use classification in Section 4.2.

Type Classification: Figure 10 shows examples of a memo and a letter. Simple functional features such as the head/body pairs in the To:, From:, and Re: fields, and the locations of the handwritten portions, allow us to distinguish between these two similar document types. Using functional features, we can achieve a gross categorization of



(a) (b)

Figure 10: Example of the differences between a memo and a letter

the documents in a database. Given a large heterogeneous database of documents, this allows us to provide groups of documents which are likely to contain some piece of requested information, even if we cannot provide the specific information. An experiment demonstrating this method of type classification will be described in Section 4.3.

Functional Enhancement: We can use the functional organization of a document to help decide which portions of it should be presented to a user and which can be ignored or considered as lower priority. The extraction of functional constructs allows this to be done without the need for content-level reasoning. In fact, many of the relationships which are explicit in the structure cannot be found at the content level; examples are the ordinal relationship between items in a list, or the spatial relationships between columns in a table. Based on these ideas, techniques can be developed to present document images to users who want to browse collections of documents. Such techniques, as illustrated in Section 4.4, make it possible to provide documents to a user in a way which is consistent with how the documents were intended to be used, or which is consistent with the goals of the reader. We believe that this will be very helpful in gaining acceptance for electronic representations of documents, since the electronic representation allows the mode of presentation of a document to be modified easily.

4 Experiments

In this section we describe some experiments on document use and type classification, and briefly outline some methods of functional enhancement. These tasks rely heavily on the identification of information units, information structures and their properties. The first step, therefore, is a segmentation of the document into appropriate information unit primitives whose properties can be used for classification or enhancement.

4.1 Extracting Information Units and Structures

In our experiments, we will consider characters, graphics blocks, and image blocks to be the basic information units. We assume that the document has been separated into text, graphics and image regions, and we then further decompose the text regions. The extraction of information units is related to the Gestalt principles, as discussed briefly in Section 2, and we rely on this in our approach to text segmentation. Proximity grouping of text is performed bottom-up to obtain a component hierarchy, and similarity grouping (boldface, italics and text size) and "good continuation" segmentation are then computed top-down.

4.1.1 Segmentation of Text

Text-based information units vary with physical scale and are dependent on the application at hand. We therefore must be able to represent multiple levels of information units. For text, the hierarchy typically consists of characters–words/phrases, lines, blocks, etc. Other units and levels are typically application-dependent–for example, strokes for handwriting, serifs for font identification, and sentences for content analysis.

Our text segmentation scheme relies on the identification of textual components by regularity (or proximity). Connected components are generated from a binary document image and the document is de-skewed using the base of each component as an indicator of its baseline. For each component, a local *proximity graph* is generated so that the relationships between a symbol and those immediately above or below it (N-S) are preserved as are relationships between a symbol and those to its left and right (E-W) (Figure 11). The symbols are then grouped appropriately. First, the dots on i's, j's, question marks, exclamation marks, etc. are identified by examining the N-S relations of a component with respect to its E-W neighbors. Next, words are created by examining the E-W regularity. The idea is that symbols in the middle of a word will be at approximately the same distance from their E and W neighbors, whereas symbols at the beginning or end of a word will be at unequal distances. Unfortunately, due to modern typesetting practices such as kerning, these distance regularities do not hold globally, and a decision about skewness must be made locally. For example, we call a symbol "W-skewed" ("E-skewed") if the distance to its west (east) neighbor in the proximity graph is greater than 1.25 times the distance to its east (west) neighbor. To handle single-character words, a symbol is not grouped with its neighbors if its E neighbor is W-skewed and its W neighbor is E-skewed. Statistical characterization of the distances in a block or line can be used to refine this process. This process can be adapted to group words into lines, lines into blocks, and blocks into columns, resulting in a hierarchical representation of the information units. Figure 12 shows line- and block-level groupings. For classification of function, the block level is sufficient; columns are only extracted for reading order.

4.1.2 Properties of Text Units

A second level of characterization is based on information unit properties. First, a gross characterization of the text height is made for each block. The height of each line's bounding box is computed, and the average height of all the lines in all multi-line blocks is computed as the average text height, based on the assumption that multi-line text blocks are a good indication of the standard "body" text of a document. Text blocks are then characterized as large or small when they vary by more then 25% from the average.

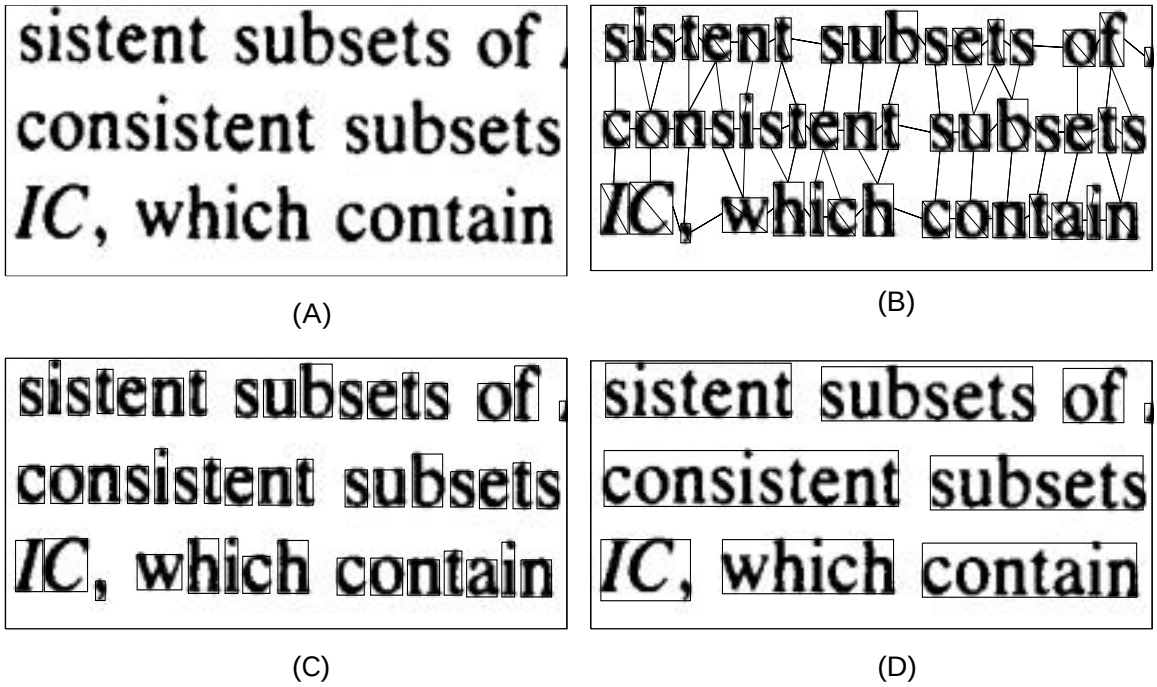


Figure 11: (A) Original image, (B) proximity graph, (C), character grouping, and (D) word grouping.

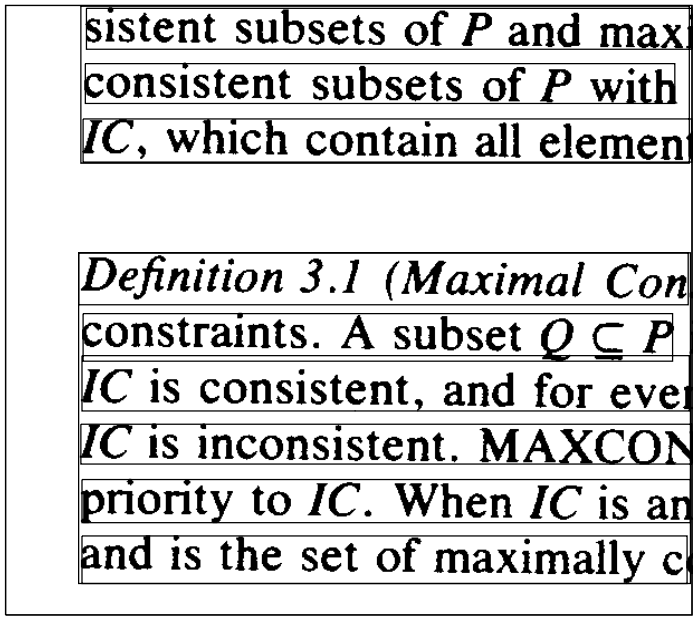


Figure 12: Line- and block-level groupings

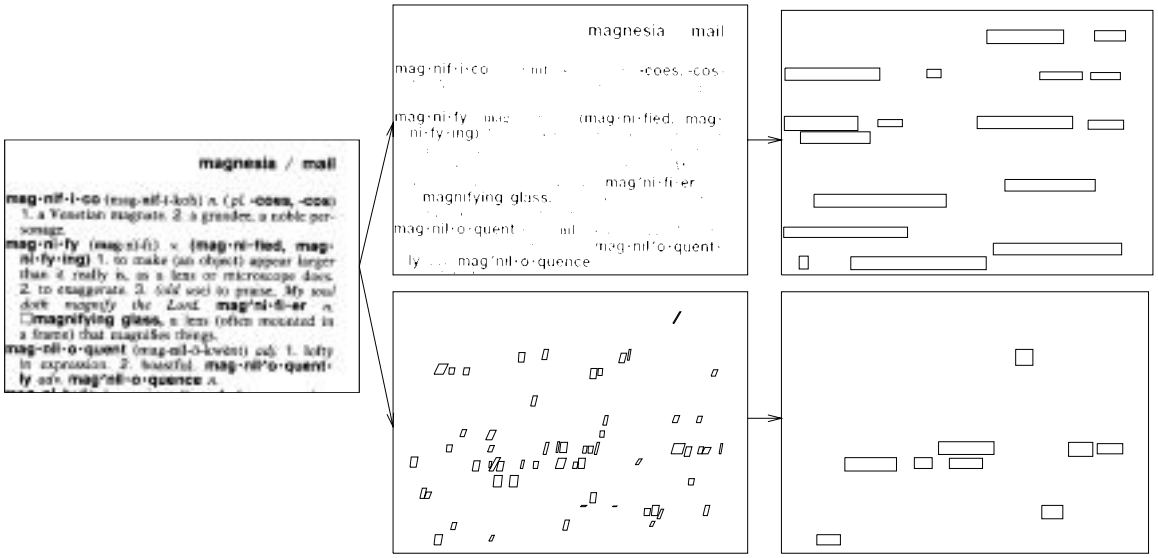


Figure 13: Boldface (top) and italic (bottom) word detection.

Words are also identified as italic or boldface. Italic words are identified by the following algorithm. The minimum upright bounding parallelogram (i.e., a parallelogram with horizontal base and top) is constructed for each component and the slant measured relative to the vertical axis. Since it is difficult to make an accurate determination of the angle from short characters, symbols taller then the average are weighted more heavily. Words in which 50% of the characters have slants greater than δ degrees are classified as italic (Figure 13). We have used $\delta = 11$ in our experiments.

Boldface is also identified at the word level, but using a morphological approach applied to individual blocks (Figure 13). An opening transform is applied in an attempt to eliminate or severely distort non-boldface text. An erosion transform is applied until more than 80% of the pixels have been eliminated, at which point a dilation is applied for an equal number of steps. When the resulting image is compared to the original image, words which are not in boldface have very limited similarity to the original while boldface character tend to remain intact. Note that boldface can be detected only in the presence of normal-weight characters, and the number of erosion steps is dependent on the scanning resolution and the size of the characters. By operating on the block level, problems caused by a wide variety of text sizes, as well as inconsistent illumination, are reduced.

4.2 Use Classification

As suggested in Section 3, the population of text blocks and their descriptions can be used to classify a document into the usage categories of reading, browsing, and searching (and modifying).

The following heuristics can be used to identify these classes:

Reading documents are characterized by a relatively small number of large text blocks on each page. The majority of the document is composed of text that has a single point size.

Browsing documents tend to have medium to large text blocks, and small text blocks of a

larger point size which act as focal points for the reader. Although readable documents have similar handles, browsable documents typically have many such handles.

Searching documents are characterized by small, repetitive text blocks.

Some of the specific properties which can be used include:

- Number of information units
- Distribution of the geometrical sizes of the units
- Number of words and lines per text block
- Geometrical arrangement of the units
- Existence of multiple point sizes
- Existence of graphic and image components

Using a set of very simple criteria, based on a subset of the above properties, we were able to classify approximately 80% of a 100-document database correctly, with approximately 5% being unclassified. The criteria used were as follows:

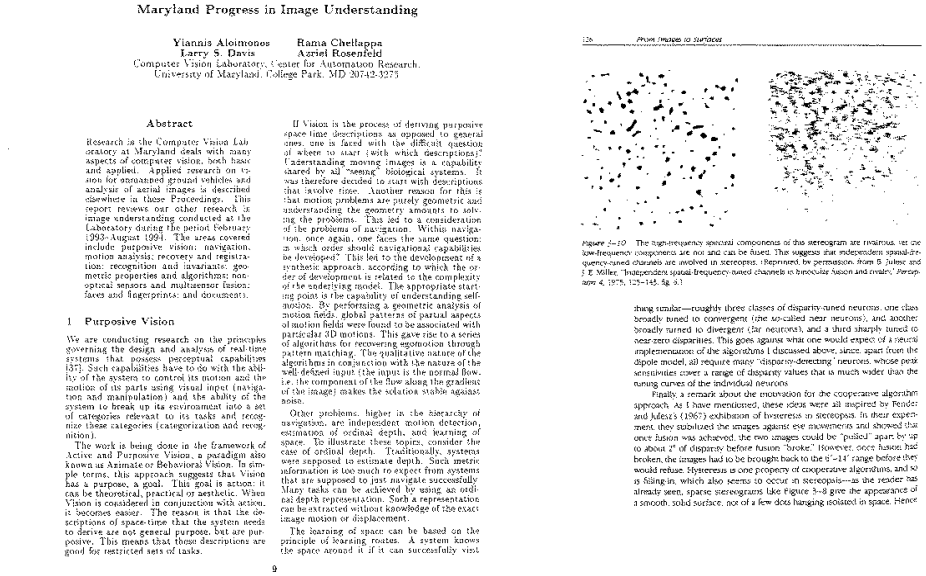
- In a searching document, no more than 25% of the text blocks should have more than five lines. There should be no image components, and few or no graphic components.
- A browsing document must have at least three head/body pairs. A head is in an emphasized font (boldface, italics, or a large font) and has no more then two lines. A body is standard text with more than two lines.
- A reading document must follow a strict (one- or two-column) column structure and must have large text blocks, primarily of a standard point size.

Figures 14-16 illustrate the use of these criteria in the block-level segmentation of reading (Figure 14), browsing (Figure 15), and searching (Figure 16) documents.

These criteria will not perform well on very complex structures. One of the difficulties is that many documents belong to more than one use class. Consider, for example, the "yellow pages" of a telephone book. The individual line listings are clearly designed for searching, but they are intermixed with "advertisements" which have browsing characteristics. Similarly, a journal article's bibliography exhibits both reading and searching characteristics.

4.3 Type Classification

Type classification is a refinement of use classification; the type of a document refers to a more specific document-level characterization such as journal article or newspaper article, or a page-level characterization such as title or contents page. We can use function-based analysis as a basis for type classification. Following Rosch [11] we regard category systems as having both vertical and horizontal dimensions. The vertical dimension concerns the level of inclusiveness (reading document → article → journal article → title page...) and



Maryland Progress in Image Understanding

Yiannis Aloimonos Rama Chellappa
Larry S. Davis Azriel Rosenfeld
Computer Vision Laboratory, Center for Automation Research,
University of Maryland, College Park, MD 20742-3275

Abstract

Research in the Computer Vision Laboratory at Maryland deals with many aspects of computer vision, both basic and applied. Applied research on vision for unmanned ground vehicles and analysis of aerial images is described elsewhere in these Proceedings. This report reviews our other research in image understanding conducted at the Laboratory during the period February 1993–August 1994. The areas covered include purposive vision; navigation, motion analysis; recovery and registration; recognition and invariants; geometric properties and algorithms; non-optical sensors and multisensor fusion; faces and fingerprints; and documents.

1 Purposive Vision

We are conducting research on the principles governing the design and analysis of real-time systems that possess perceptual capabilities [37]. Such capabilities have to do with the ability of the system to control its motion and the motion of its parts using visual input (navigation and manipulation) and the ability of the system to break up its environment into a set of categories relevant to its tasks and recognize these categories (categorization and recognition).

The work is being done in the framework of Active and Purposive Vision, a paradigm also known as Animate or Behavioral Vision. In simple terms, this approach suggests that Vision has a purpose, a goal. This goal is action: it can be theoretical, practical or aesthetic. When Vision is considered in conjunction with action, it becomes easier. The reason is that the descriptions of space-time that the system needs to derive are not general purpose, but are purposive. This means that those descriptions are good for restricted sets of tasks.

If Vision is the process of deriving purposive space time descriptions as opposed to general ones, one is faced with the difficult question of where to start (with which descriptions)? Understanding moving images is a capability shared by all "seeing" biological systems. It was therefore decided to start with descriptions that involve time. Another reason for this is that motion problems are purely geometric and understanding the geometry amounts to solving the problems. This led to a consideration of the problems of navigation. Within navigation, once again, one faces the same question: in which order should navigational capabilities be developed? This led to the development of a synthetic approach, according to which the order of development is related to the complexity of the underlying model. The appropriate starting point is the capability of understanding self-motion. By performing a geometric analysis of motion fields, global patterns of partial aspects of motion fields were found to be associated with particular 3D motions. This gave rise to a series of algorithms for recovering egomotion through pattern matching. The qualitative nature of the algorithms in conjunction with the nature of the well-defined input (the input is the normal flow, i.e. the component of the flow along the gradient of the image) makes the solution stable against noise.

Other problems, higher in the hierarchy of navigation, are independent motion detection, estimation of ordinal depth, and learning of space. To illustrate these topics, consider the case of ordinal depth. Traditionally, systems were supposed to estimate depth. Such metric information is too much to expect from systems that are supposed to just navigate successfully. Many tasks can be achieved by using an ordinal depth representation. Such a representation can be extracted without knowledge of the exact image motion or displacement.

The learning of space can be based on the principle of learning routes. A system knows the space around it if it can successfully visit

9

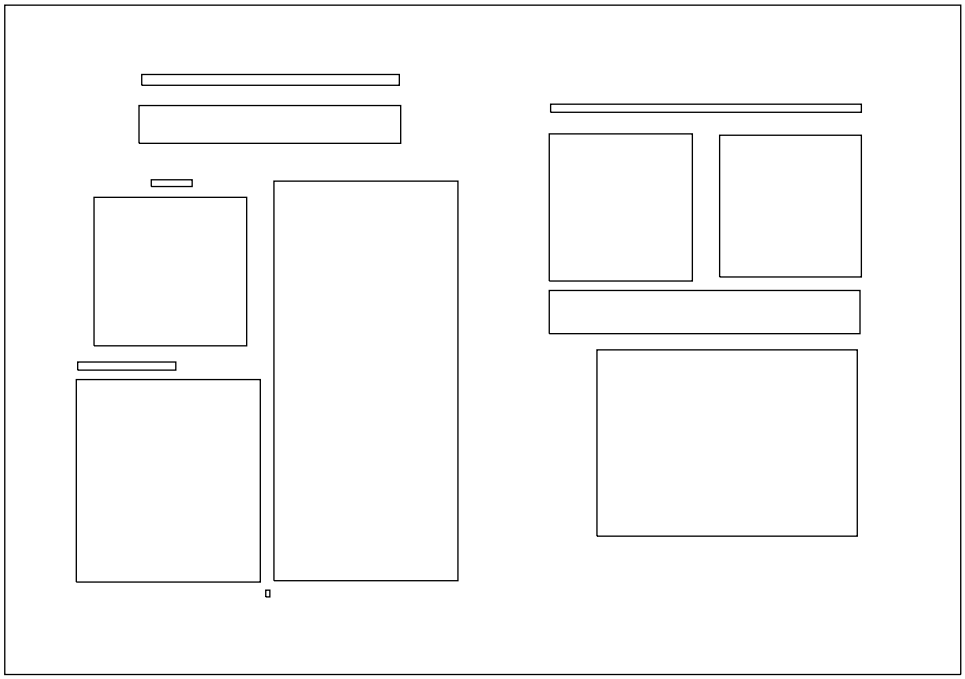
126 From Images to Surfaces

Figure 5–10 The high-frequency spectral components of this stereogram are irrelevant, yet the low-frequency components are not and can be fused. This suggests that independent spatial-frequency-tuned channels are involved in stereopsis. (Reprinted, by permission, from B. Julesz and J. E. Miller, "Independent spatial-frequency-tuned channels in binocular fusion and rivalry," *Perception* 4, 1975, 125–143, fig. 6.)

thing similar—roughly three classes of disparity-tuned neurons, one class broadly tuned to convergent (the so-called near neurons), and another broadly tuned to divergent (far neurons), and a third sharply tuned to near-zero disparities. This goes against what one would expect of a neural implementation of the algorithms I discussed above, since, apart from the dipole model, all require many 'disparity-detecting' neurons, whose peak sensitivities cover a range of disparity values that is much wider than the tuning curves of the individual neurons.

Finally, a remark about the motivation for the cooperative algorithm approach. As I have mentioned, these ideas were all inspired by Julesz and Julesz's (1967) exhibition of hysteresis in stereopsis. In their experiment, they stabilized the images against eye movements and showed that once fusion was achieved, the two images could be "pulled" apart by up to about 2° of disparity before fusion "broke." However, once fusion had broken, the images had to be brought back to the 0'–14' range before they would refuse. Hysteresis is one property of cooperative algorithms, and so is filling-in, which also seems to occur in stereopsis—as the reader has already seen, sparse stereograms like Figure 5–8 give the appearance of a smooth, solid surface, not of a few dots hanging isolated in space. Hence

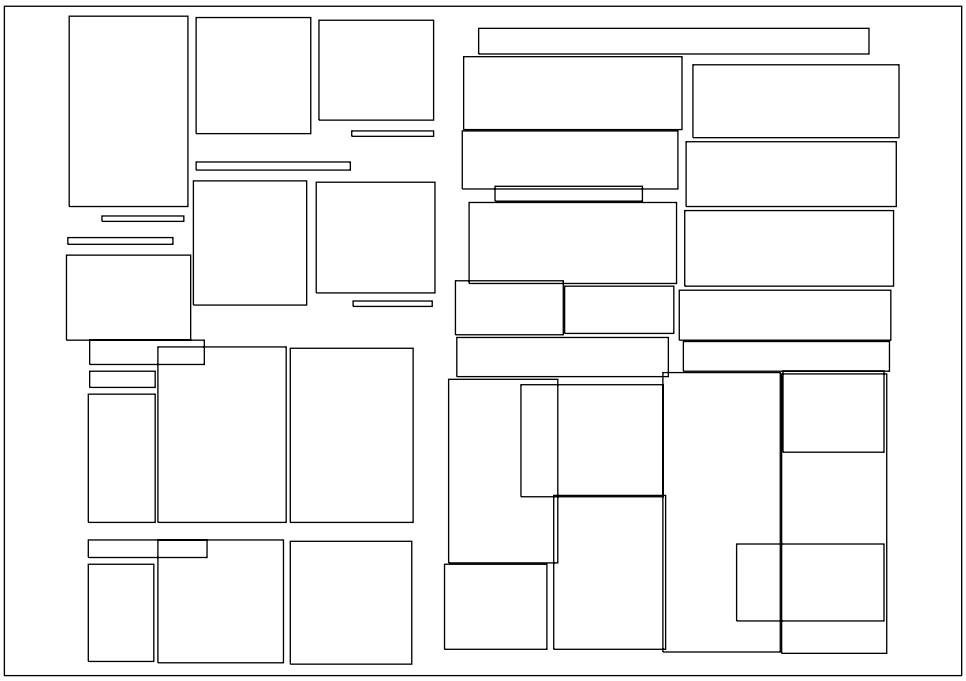
(a)



(b)

Figure 14: Reading document segmentation

(a)



(b)

Figure 15: Browsing document segmentation

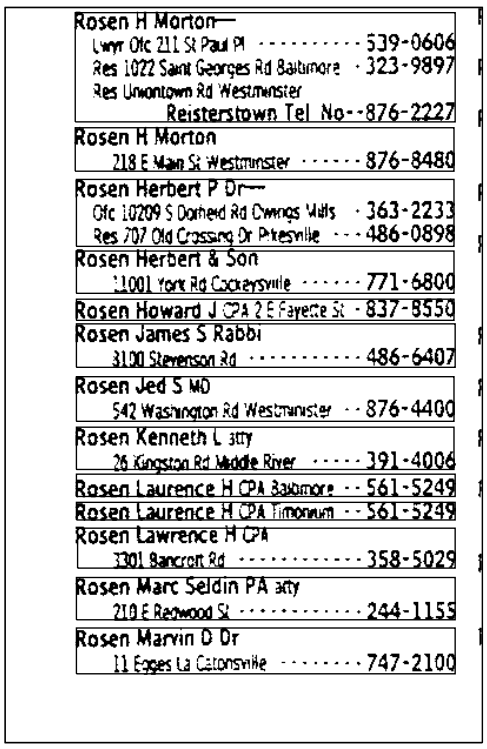


Figure 16: Searching document segmentation

the horizontal dimension concerns classes at the same level of inclusiveness (the dimension on which a newspaper, a novel and a phone book vary, for example).

Using this terminology, we can classify documents starting from a superordinate (high) level and moving down to subordinate levels using function as the discriminating property. The elements which constitute a document have different functionalities. Their geometries are loosely constrained by the need to fulfill these functions. For example, in a newspaper, components such as headlines, headers, columns and figures all support different functions. Their combination defines the document's functionality which is a basis for document classification. Using this approach provides us with the power of functional recognition. A small knowledge base suffices to type-classify a wide variety of documents.

Taking the same approach as described in [12, 13, 14], we can treat our system's knowledge as a frame system organized into a tree structure, as illustrated in Figure 17. The root node represents a superordinate category (document: reading), and the immediate children of the root represent basic level categories (article and novel). The categorization can be performed by identification of functional elements in the configuration by associating them with their functional labels. Checking if a document can serve as an X (e.g. a journal article) involves deciding whether the proper functional requirements are met. This is done using the same mechanism of "knowledge primitives" (KPs) as used in [12, 14]. A KP is a type of parameterized procedure call which makes low-level observations about a document. For example, we can use a KP of the form info_unit(document_element, info_unit_type, range_parameters). This KP can be used to determine if the width, length or size of an information unit lies within a specified range. Combining a number of KPs provides a categorization capability.

The classification process can use the tree structure as a control structure. A category

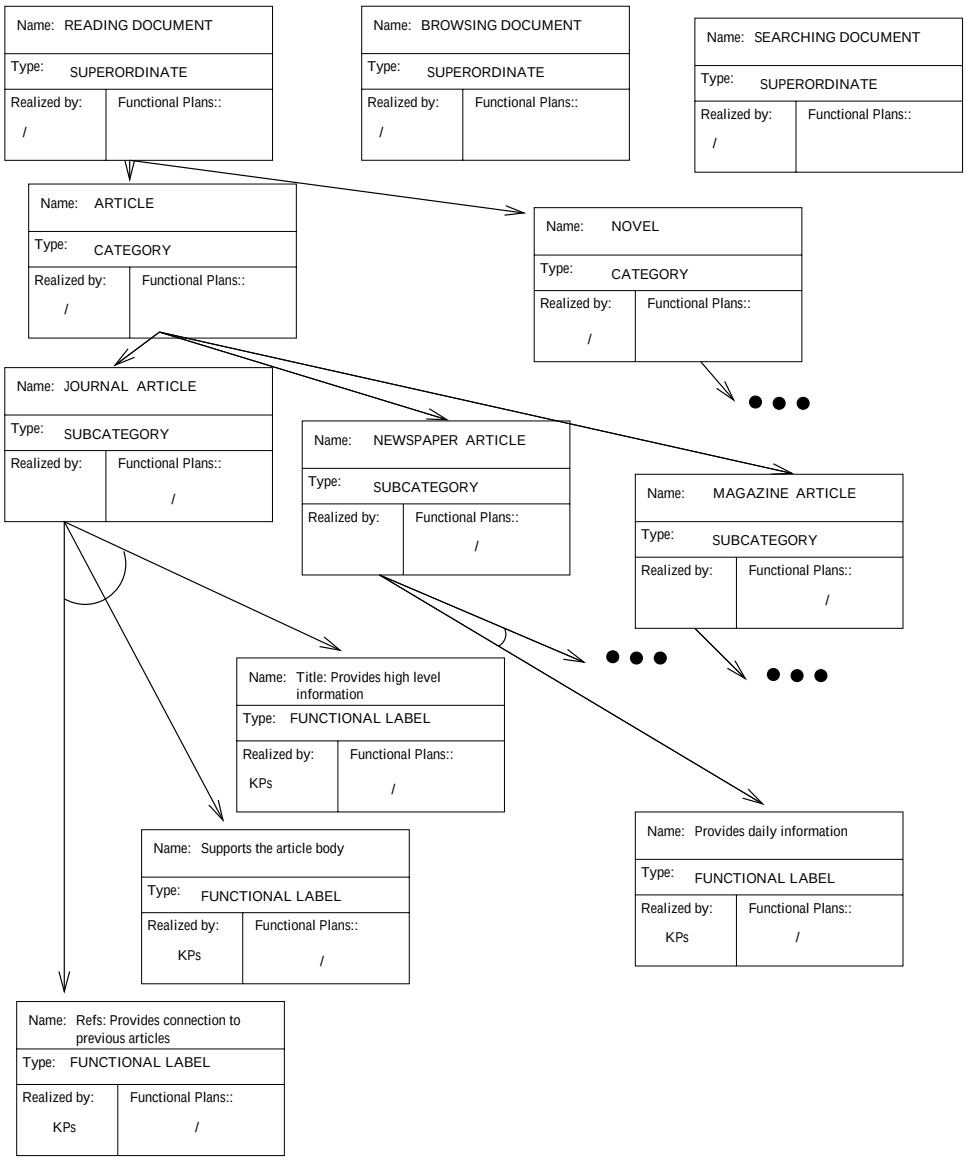


Figure 17: A partial category tree for reading documents.

can be hypothesized (see [14] for more details), or given by some top-level program. Once a category is selected for analysis, the subtree of the category is used to activate appropriate KPs. As the traversal of the tree proceeds, the system attempts to categorize the input document as belonging to some sub-category by confirming that all the functional requirements are met.

In the next section, we describe and provide experimental results for an approach to "learning" a set of KPs that can categorize journal article pages.

4.3.1 Classifying Journal Pages

A first level of inclusiveness below "journal article" is the level of individual pages. As an example of how to perform classification at this level, we ran a set of experiments using the George Mason University AQ15c rule learning system [17]. The goal was to classify individual journal pages as being title, reference or body.

A set of 59 journal page images from the University of Washington English Document Image Database-I was used for training and testing. This database contains images of pages as well as page- and zone-level ground truth for each page. Each description includes general characteristics of the page and characteristics of each zone on the page. The page characteristics include, for example, "dominant-font-size", "dominant-font-style", and "number-of-columns", while the zone characteristics include, for example, "type", "location", "text-alignment", and "dominant-font-style". The classification of pages into the three categories was not provided in the ground truth, and was performed manually.

For our experiments we used a subset of the page characteristics. We also defined some additional attributes by agglomerating the original attribute values. These new attributes were selected in such a way that they could be automatically derived from the database images.

The complete database was converted to Document Interchange Format (DIF). In this format, each page is described by specifying general information about the page (records labeled PAGE), and a list of zone descriptions (records labeled ZONE).

Figure 18 shows an example of a page; its zones are described below:

```
PAGE,read-A00G,normal,plain,1
ZONE,000,text,2288 244 2344 288,justified,normal,plain,0
ZONE,001,text,768 1548 2240 1628,justified,normal,plain,1
ZONE,002,text,760 1660 2324 2108,justified,normal,plain,1
ZONE,003,text,756 2208 968 2260,justified,normal,emphasis,1
ZONE,004,text,752 2312 2320 2564,justified,normal,plain,1
ZONE,005,graphics,956 296 2264 1472,non-text,non-text,non-text,0
```

We constructed a representation space for learning by starting with a fixed set of attributes, and automatically determining sets of attribute values which sufficed to classify the training set. Some of the attributes used to create the representation space are given in Table 1. Note that only structural attributes are employed; no content information is used.

4.3.2 Rule Learning

The set of 59 pages was split into two sets, one set for training the learning algorithm and the second set for testing prediction accuracy. The AQ15c system was used for learning

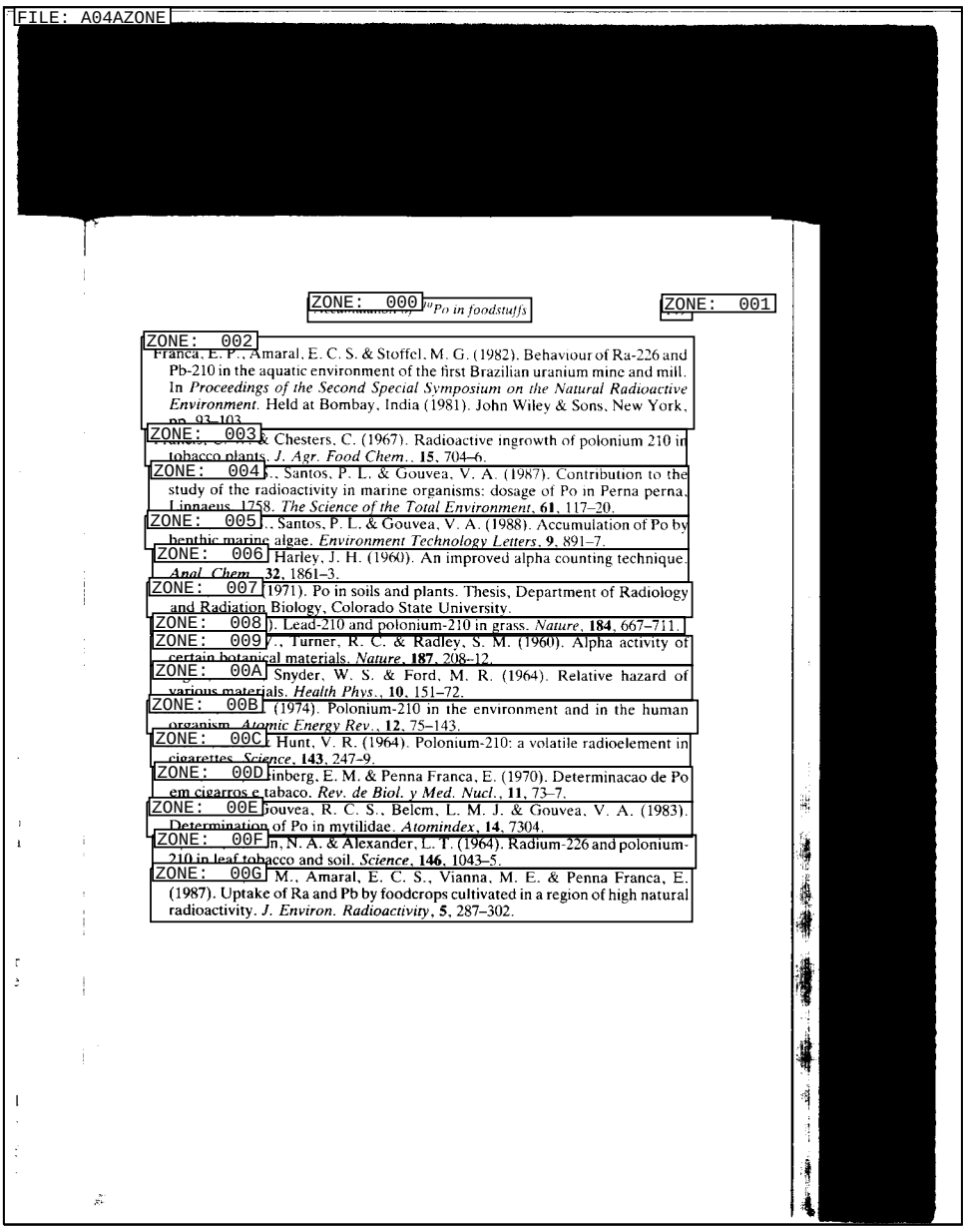


Figure 18: An example page and its zones

ID	Name	Description
1	tz00	Number of zones in left-top section
2	tz01	Number zones in left-mid section
3	tz02	Number of zones in left-bot section
4	tz10	Number of zones in right-top section
5	tz11	Number of zones in right-mid section
6	tz12	Number of zones in right-bot section
7	pDFSz	Dominant font size
8	pDFSt	Dominant font style
9	pDZA	Dominant zone alignment
10	pC	Number of columns
11	pTZ	Number of text zones
12	pGZ	Number of graphic zones
13	pIZ	Number of image zones
14	pRZ	Number of ruling zones
15	azs	Average zone size
16	hVZ	1 if header has variable length zones, 0 otherwise
17	hZS	1 if average zone size in header area > 4, 0 otherwise
18	sMZ	Maximum number of consecutive zones with similar height/width

Table 1: Representation space

	Range		
Attribute	Reference	Title	Body
sMZ	[1,7]	[1,3]	[1,2]
azs	[1,4]	[3,7]	[4,19]

Table 2: Rules generated by the AQ15c system

classification rules. The rules generated by the system could vary depending on a number of control parameters.

The goal was to produce a (preferably) small set of rules which could be used to distinguish between the three classes. The rules derived by the learning system for reference, title and body pages were consistent with the functional descriptors described previously. In particular, the most discriminatory attributes turned out to be the number of vertically neighboring zones with consistent height (sMZ) and the average size of the zones (azs). These attributes had different ranges for pages belonging to the three classes as illustrated in Table 2.

4.3.3 Results and Discussion

We used 38 of the 59 documents for training. Using the resulting rules we were able to obtain 100% classification accuracy for the training set and over 90% for the testing set as shown in Table 3. The rules are intuitively plausible and highly consistent with our functional principles. The number and average size of the information units (zones) play major roles in the rules.

Figure 19: Pages classified as body (top), reference (middle) and title (bottom).

Examples of documents that were classified into each class are shown in Figure 19. Note that the second example of a reference page is also a title page.

4.4 Functional Enhancement

If we can decompose a document into functional components, we can use its functional organization to help decide which portions of it should be presented to the user and which can be ignored or considered as lower priority. The extraction of functional constructs allows this to be done without the need for content-level reasoning. Using these ideas,we can present document images to users in accordance with their goals. If a user wants, for example, to browse collections of documents, we can provide only the upper-level headers, and give the user the option to retrieve full information when needed.

The pieces of a document which we choose to provide are based on the observation that

	Training			Testing		
Type	Num Samples	Num Correct	Accuracy	Num Samples	Num Correct	Accuracy
Title	12	12	100%	7	6	86%
Reference	12	12	100%	7	6	86%
Body	14	14	100%	7	7	100%

Table 3: Type classification results

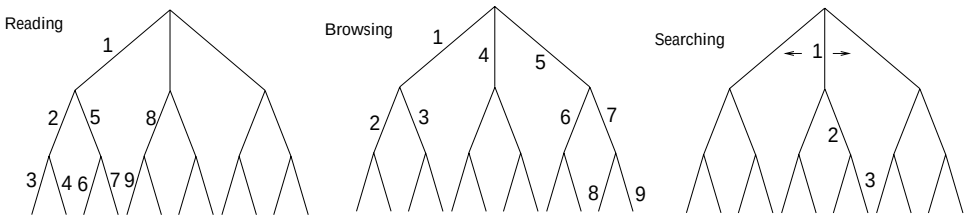


Figure 20: Examples of navigational trees associated with reading, browsing and searching

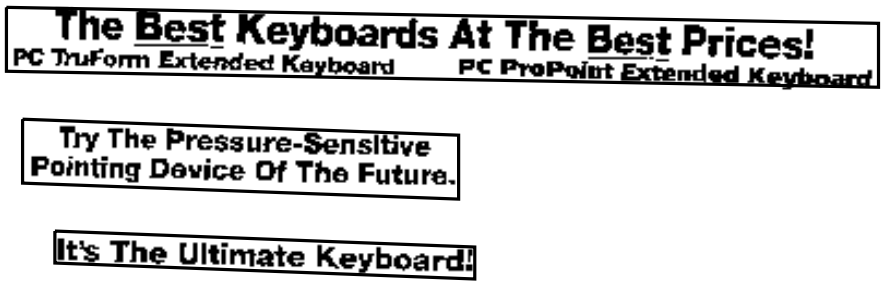


Figure 21: Enhanced browsing capability

there appears to be a close analogy between these three modes of document usage and three methods of traversal of a tree structure (Figure 20). Reading a document corresponds to a depth-first search of the tree. We expand each node in turn and traverse the tree depth-first. Browsing resembles a pruned depth-first search; the reader identifies nodes at higher levels which are of interest, and prunes those which are not. Searching can be implemented by treating the tree as a decision tree; a node or set of nodes is explored at each level, until the one which contains the appropriate information is found, and a decision is made as to which node to explore further. Backtracking is typically limited, but can easily be provided when errors are made. We use these ideas in the following examples.

Assume that a user wants to browse through a document which consists of pages like the one presented in Figure 15(a). We can present the information in a manner consistent with the traversal mode by giving the title of each information unit (see Figure 21), and allowing the user to ask for the full unit if needed.

For searching a document, we can present the beginning of each information unit, yielding a compressed representation that allows for acceleration in the decision process. For example, the search document shown in Figure 16 can be presented in a compressed form such as shown in Figure 22. (More generally, in an alphabetically organized search document, only the first few characters on a page need be presented at the highest level, and the first few characters of each listing at a lower level.)

These examples also demonstrate the usefulness of the electronic representation of documents, since this representation allows the mode of presentation of a document to be modified easily, according to the user's goals and needs.

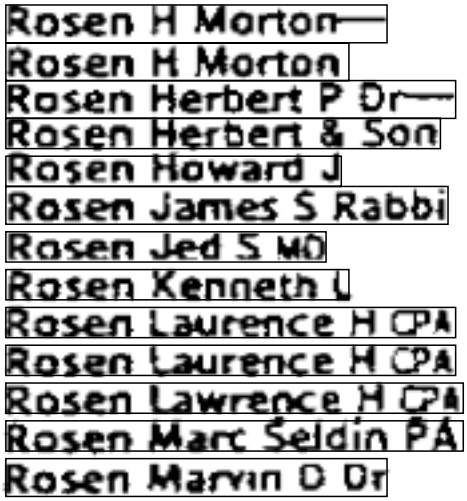


Figure 22: Enhanced search capability

5 Discussion and Conclusions

Document functionality relates to how the document conveys information to its user. In this paper, we have provided a basis for understanding the functional aspects of document design and usage. Authors use layout and emphasis to make it easier to extract information from documents. Traditional document understanding and conversion techniques have ignored the intended functionality of the document, especially its class-independent functional structure. An important advantage of our approach is that it provides an ability to organize documents without understanding their content.

We plan to extend our work to provide a more complete taxonomy of functional primitives, and to implement a full-scale system for functional typing and document classification.

References

[1] H.S. Baird. Anatomy of a versatile page reader. *Proceedings of the IEEE*, 80:1059–1065, 1992.

[2] H.S. Baird, H. Bunke, and K. Yamamoto. *Structured Document Image Analysis.* Springer, Berlin, 1992.

[3] K.E. Bullen. *An Introduction to the Theory of Machines.* Cambridge University Press, Cambridge, UK, 1971.

[4] A. Dengel, R. Bleisinger, F. Fein, R. Hoch, F. Hones, and M. Malburg. OfficeMAID — A system for office mail analysis, interpretation and delivery. In *International Workshop on Document Analysis Systems*, pages 253 – 276, 1994.

[5] International Standard Organisation. *Text and Office Systems—Office Document Architecture (ODA) and Interchange Format*, 1989. International Standard 8613.

[6] K. Koffka. *Principles of Gestalt Psychology.* Harcourt, Brace & World, New York, 1935.

[7] M. Krishnamoorthy, G. Nagy, S. Seth, and M. Viswanathan. Syntactic segmentation and labeling of digitized pages from technical journals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:737–747, 1993.

[8] L. O'Gorman. The document spectrum for page layout analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:1162 – 1173, 1993.

[9] E. Rivlin, S.J. Dickinson, and A. Rosenfeld. Recognition by functional parts. *Computer Vision and Image Understanding*, 62:164–177, 1995.

[10] E. Rivlin and A. Rosenfeld. Navigational functionalities. *Computer Vision and Image Understanding*, 62:232–247, 1995.

[11] E. Rosch. *Cognition and Categorization.* Erlbaum, Hillsdale, NJ, 1978.

[12] L. Stark and K. Bowyer. Achieving generalized object recognition through reasoning about association of function to structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:1097–1104, 1991.

[13] L. Stark and K. Bowyer. Indexing function-based categories for generic object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 795–797, 1992.

[14] L. Stark and K.W. Bowyer. Function-based generic recognition for multiple object categories. *CVGIP: Image Understanding*, 59:1–21, 1994.

[15] S.L. Taylor. Information-based document analysis systems in a distributed environment. In *International Workshop on Document Analysis Systems*, pages 93 – 108, 1994.

[16] T. Watanabe, Q. Luo, and N.Sugie. Structure recognition methods for various types of documents. *Machine Vision and Applications*, 6:163–176, 1993.

[17] J. Wnek, K. Kaufman, E. Bloedorn, and R.S. Michalski. Inductive learning system AQ15c: The method and user's guide. Technical Report MLI 95-4, Machine Learning and Inference Laboratory, George Mason University, March 1995.